

Применение системы лексических значений предикатных слов для автоматизации пополнения словаря Моделей Управления.

Д.В.Михайлов, Г.М.Емельянов

Новгородский государственный университет имени Ярослава Мудрого

Настоящая работа посвящена (*плакат 1*) решению проблемы автоматизации пополнения языковой Базы Знаний в задаче установления смысловой (семантической) эквивалентности высказываний Естественного Языка (ЕЯ).

К числу наиболее актуальных из указанных задач относится интерпретация результатов открытых тестов в системах компьютерного дистанционного обучения и контроля знаний. Тестовые задания открытой формы требуют от обучаемого формулирования развернутого ответа на поставленный системой вопрос (*плакат 2*). Как показывает опыт разработки различных тестовых систем, применение открытых тестов затруднено в силу ряда причин. Одна из них заключается в необходимости оперирования большим количеством сущностей при интерпретации теста и, как следствие, отсутствии универсальных механизмов оценки правильности ответа. Наиболее разумным путем решения указанной проблемы является введение в рассмотрение “эталонного” смысла, относительно которого ведется сравнение (*плакаты 3 и 4*). Причем эталонный ответ также формулируется на естественном языке самим преподавателем-разработчиком теста. В этом случае обработка результата теста сводится к сравнению смыслов двух высказываний естественного языка : ответа, задуманного преподавателем как правильный (эталонный), и ответа, введенного обучаемым.

Применение языка глубинного синтаксиса в качестве языка смыслов в рамках теоретического подхода к языку как преобразователю “Смысл \Leftrightarrow Текст” дает возможность использования для сравнения смыслов высказываний конечного числа корректно формализуемых преобразований помеченных деревьев (*плакат 6*). Основным достоинством такого подхода является независимость правил синонимических преобразований от предметной области высказываний. Указанные правила синонимических преобразований описывают ситуации лексико-синтаксических замен на уровне варьирования универсальной (абстрактной) лексикой в рамках аппарата стандартных Лексических Функций (ЛФ), что особенно актуально для реальных тестов : в большинстве случаев обучаемый употребляет синонимы именно на уровне абстрактных слов и их сочетаний, оставляя предметную лексику без изменений (*плакаты 5 и 7*).

Построение совокупности деревьев глубинного синтаксиса фраз анализируемого высказывания требует последовательного выполнения морфологического, синтаксического анализа и нормализации полученных деревьев синтаксического подчинения – преобразование в глубинные синтаксические структуры (*плакат 8*). При этом для идентификации типов отношения подчинения между лексемой и ее глубинными синтаксическими

актантами используется информация Моделей Управления (МУ, плакаты 9 и 10¹).

Действующая система анализа смысловой эквивалентности ЕЯ-текстов, построенная на базе указанного теоретического подхода, должна иметь в своем составе описание МУ всех предикативных лексем используемого подмножества ЕЯ. Данное требование относится как к универсальной (абстрактной) лексике, за счет которой обеспечивается смысловое варьирование, так и к предметно-ориентированной лексике. Описания МУ абстрактных лексем могут быть заложены в языковую базу знаний изначально, при проектировании системы. Для предметных лексем их МУ описываются при настройке системы на конкретную область знаний (область знаний, по которой предполагается строить тесты – в примере с дистанционным обучением).

Описание МУ вручную “с нуля” требует от настройщика системы определенных лингвистических навыков, владения подходом “Смысл \Leftrightarrow Текст” в совершенстве, что в практике построения, в частности, систем тестирования знаний не представляется возможным.

К настоящему моменту необходимая для машинного анализа текстов семантическая информация лексем русского языка (включая МУ) наиболее полно отражена в Русском общесемантическом словаре (РОСС). Идеология РОСС имеет практическое воплощение в разработанном рабочей группой Aot.ru АРМ лингвиста (<http://www.aot.ru>). Тем не менее, при описании МУ новых слов *актуальна проблема* адекватной идентификация ролевых ориентаций семантических валентностей, поскольку ролевая ориентация валентности даже в упомянутом АРМ задается самим настройщиком, вручную.

Исходя из вышеизложенного, *цель* настоящей работы сформулирована как (плакат 1) *исследование возможностей использования существующих в языке закономерностей для решения проблемы автоматизации построения МУ и контроля их корректности.*

Наиболее естественный путь решения проблемы автоматизации построения МУ новых слов вытекает из показанного академиком Ю.Д. Апресяном соответствия между толкованием на ЕЯ лексического значения предикатного слова (лексикографическим толкованием) и его МУ. В теоретической лексикографии с лексическим значением слова связывается предмет или явление действительности, которые обозначены этим словом. Как показал Ю.Д. Апресян, состав семантических валентностей слова определяется анализом обозначаемой им ситуации. Семантические валентности вытекают непосредственно из лексического значения слова а, следовательно, из его толкования.

¹ В соответствии с предложенным И.А.Мельчуком и А.К.Жолковским строением словарной статьи Толково-комбинаторного словаря, отношение между лексемой и ее глубинным синтаксическим актантами по МУ в настоящей работе описывается представленной на *плакатах 9 и 10* структурой в виде составных объектов языка Пролог.

На основе полученных Ю.Д. Апресяном теоретических выводов были сформулированы следующие задачи исследования :

- 1) Выделить составляющие (элементы) толкования слова, которые соответствуют описанию отдельного актанта и могут задаваться настройщиком;
- 2) Исследовать зависимости между составляющими толкования и элементами описания семантических валентностей слова;
- 3) Выявить существующие зависимости между отдельными составляющими описания глубинного актанта слова с целью выделения ключевых элементов в соответствии с предложенной И.А. Мельчуком концепцией МУ;
- 4) На основе выделенных ключевых элементов описания актанта слова разработать методику систематизации и контроля корректности информации справочника МУ;
- 5) В соответствии с выявленными зависимостями разработать алгоритмическую составляющую перехода от фрагмента толкования слова к описанию его глубинного синтаксического актанта;
- 6) Разработать структуру базы знаний программного комплекса ведения справочника МУ.

Для решения первой и второй из поставленных задач на приведенных Ю.Д. Апресяном примерах был последовательно рассмотрен процесс описания семантической валентности слова на основе фрагмента толкования. При этом в качестве исходных данных при описании отдельного актанта выделены (*плакат 11*) : вопрос для заполняющей формы (вопросительное местоимение+предлог) и слова для обозначения семантической ориентации актантов. Путем использования указанных составляющих толкования последнее может быть представлено в виде дерева на *рис.2 (плакат 11)*. Вопрос для заполняющей формы позволяет определить морфологический способ (*плакат 9, 12*) реализации будущей валентности.

Как показано Ю.Д. Апресяном, при описании лексикографического толкования на естественном языке неявным образом задается вполне определенный порядок появления в тексте словоформ, реализующих ту или иную валентность. С учетом показанных И.А. Мельчуком ограничений на число актантов, связанных с предикатным словом отношением подчинения заданного типа в настоящей работе предложено представленное на *плакате 13* правило установления типа отношения глубинного синтаксиса (D_synt, *плакат 9*). Распознанный тип отношения глубинного синтаксиса может быть использован для идентификации синтаксического класса и грамматических характеристик актанта в случае сходства лексики вопросов для различных грамматических форм. Для этой цели достаточные для взаимного различия типичные морфологические формы выражения валентностей выносятся в отдельную таблицу (*таблица 2, плакат 13*).

Авторами настоящей работы для решения задачи автоматического распознавания ролевой ориентации семантической валентности были сформулированы представленные на *плакате 14* свойства семантических

валентностей, которые составляют основу представленной на *плакате 15* их классификации и алгоритма распознавания. Этот алгоритм использует представленную в табличной форме информацию о возможных способах реализации ролевых зависимостей в качестве семантических валентностей (*плакаты 16, 17 и 18*). Будучи реализованным в среде Visual Prolog 5.2 с представленной на *плакате-приложении 2* структурой программного комплекса, предложенный алгоритм в общем случае выдает несколько возможных ролевых ориентаций. Каждому варианту ролевой ориентации соответствует свой вариант МУ предикатного слова, который отражает тот или иной смысловой оттенок.

Лексическое значение (семантический класс) предикатного слова определяется анализом обозначаемой им ситуации и, следовательно, может быть охарактеризовано набором актантов этой ситуации. Причем каждый актант соответствует некоторой семантической валентности из описываемых посредством МУ. Каждая валентность характеризуется ролевой ориентацией и семантическим содержанием. Последнее может быть выражено перечислением семантических классов слов, способных замещать данную валентность. Таким образом, ролевая ориентация и семантическое содержание выступают в качестве ключевых атрибутов актанта нового слова. Сравнение ролевого состава лексического значения добавляемого слова и ранее описанных слов позволяет выявить между ними отношения “субконцепт-суперконцепт”. Представление совокупности лексических значений предикатных слов в форме показанного на *плакате 20* частного случая многозначного лексикографического контекста позволяет систематизировать лексические значения предикатных слов используемого подмножества ЕЯ и тем самым определить место нового слова в существующей лексической системе.

Отношение “субконцепт-суперконцепт” между отдельными лексическими значениями в простейшем случае будет иметь место при наличии показанного на *плакате 21* взаимно-однозначного соответствия между множествами семантических классов актантов с идентичными ролевыми ориентациями у гипонима и гиперонима. Тем не менее, при наличии у гипонима семантических валентностей, производных от валентностей гиперонима, возникает проблема безошибочной идентификации указанного отношения. Примером таких валентностей может послужить валентность получателя (*Recip*) у глаголов Семантического Класа “*передача в распоряжение*” и образованная от нее валентность контрагента (*Contrag*) у предикатов со значением “*купи-продажи*”, *плакаты 22 и 24*. Для случая производных валентностей авторами настоящей работы предлагается задействовать (*плакат 22*) информацию дескрипторов таксономических категорий (КАТ) и Семантических Характеристик (СХ)² из

² В приведенном на *плакате 21* примере описаний семантических классов слов верхней окрестности глагола *арендовать* используются семантические характеристики со следующими значениями: “FINANCIAL” – означает все, что связано в той или иной мере с финансами; “CAUS” – оператор каузации; “BELNG” – собственность, владение; “LEGISL” – все, что связано с какими-либо нормативными актами,

используемых в словаре РОСС. При этом взаимно-однозначное соответствие между семантическими классами актанта гипонима и гиперонима устанавливается путем поиска списка семантических характеристик одного из них в качестве подсписка у другого. Списки семантических характеристик лексических значений для предикатных слов находятся в задаваемом *определением 2, плакат 23* соответствии.

Путем проверки наличия отношения “субконцепт-суперконцепт” между различными парами лексических значений может быть построена концептуальная решетка, из которой на экране отображается окрестность лексического значения нового слова. Средства Visual Programming Interface (VPI) позволяют организовать вывод на экран ролевого состава лексических значений из древовидного графического представления окрестности добавляемого слова, а также собственные окрестности этих лексических значений. На плакате 24 представлена верхняя окрестность лексического значения глагола “арендовать”.

В целях оценки адекватности используемых при описании семантических классов наборов дескрипторов (прежде всего – Семантических Характеристик) в настоящей работе путем применения реализующего методы ФКА специализированного ПО ToscanaJ (<http://toscanaj.sourceforge.net/>) строится представленная на *плакате 25* модель системы Лексических Значений предикатных слов. При этом Лексические Значения выступают в качестве объектов, а Семантические Характеристики – в качестве атрибутов.

В *перспективе* появляется возможность использования построенной описанным путем системы лексических значений предикатных слов для подбора элементов лексикографического толкования, в наибольшей степени соответствующих описанию актантов нового слова и, тем самым, сократить объем перебора, который имеет место при анализе представленных на естественном языке фрагментов толкования.

В качестве *перспективного направления исследований* следует также отметить применение предложенного А.С. Нариньяни семантически-ориентированного подхода³ для анализа фрагментов толкования и выделения значимой для построения МУ информации (типичной лексики обозначения ролей, семантической ориентации актантов, типов отношений глубинного синтаксиса между словом и актантом). Здесь необходимо задействовать знания о лексикографическом толковании как жанре Естественного Языка.

законодательством; с помощью “PERIO”, “TIME” задается смысловой оттенок временного периода, интервала. Слова семантических классов из приведенных на *плакате 22* примеров относятся к таксономической категории слов-этикеток (LABL), обозначающих ситуации (SIT).

³ Суть семантически-ориентированного подхода в том, что лексическим единицам языка (лексемам, словокомплексам) приписываются определенные семантические классы, выражающие смысл данной лексической единицы в рассматриваемом регистре (жанре) ЕЯ. Кроме того, некоторые семантические классы имеют в качестве атрибута семантическую ориентацию, которая в каждом конкретном случае связывает данное слово с определенным элементом модели предметной области.